Contribution of Color Information in Visual Saliency Model for Videos

Shahrbanoo Hamel, Nathalie Guyader, Denis Pellerin, Dominique Houzet

Contribution Of Color Information In Visual Saliency Model For Videos

Shahrbanoo Hamel, Nathalie Guyader, Denis Pellerin, and Dominique Houzet

GIPSA-lab, UMR 5216, Grenoble, France*

Abstract. Much research has been concerned with the contribution of the low level features of a visual scene to the deployment of visual attention. Bottom-up saliency models have been developed to predict the location of gaze according to these features. So far, color besides to brightness, contrast and motion is considered as one of the primary features in computing bottom-up saliency. However, its contribution in guiding eye movements when viewing natural scenes has been debated. We investigated the contribution of color information in a bottom-up visual saliency model. The model efficiency was tested using the experimental data obtained on 45 observers who were eye tracked while freely exploring a large data set of color and grayscale videos. The two datasets of recorded eye positions, for grayscale and color videos, were compared with a luminance-based saliency model [1]. We incorporated chrominance information to the model. Results show that color information improves the performance of the saliency model in predicting eye positions.

Keywords: color information, visual saliency, video, eye tracking

1 Introduction

The mechanism of visual attention allows selecting the relevant parts of a visual scene at the very beginning of exploration. The selection is driven by the properties of the visual stimulus through bottom-up processes, as well as by the goal of observer through top-down processes [2], [3]. Visual attention models tend to predict the parts of the scene that are likely to deploy the attention [4], [5], [6], [1]. Most of the models are bottom-up models based on the Feature Integration and Guided Search theories [7], [8]. These theories stipulate that some elementary salient visual features such as intensity, color, depth and motion, are processed in parallel at a pre-attentive stage, subsequently combined to drive the focus of attention. This approach is in accordance with the physiology of the visual system. Hence, in almost all the models of visual attention, low level features like intensity, color, spatial frequency

* This research was supported by Rhone-Alpes region (France) under the CIBLE project No. 2136. Thanks to D. Alleysson and D. Meary for providing us with spectrometer measurements.

are considered to determine the visual saliency of regions in static images, whereas motion and flicker are also considered in the case of dynamic scenes [4], [6], [1]. More recently, the contribution of different features like color in guiding eye movements when viewing natural scenes has been debated. Some studies suggested that color has little effect on fixation locations [9], [10], [11], which brings to question the necessity of the inclusion of color features in the saliency models [12]. In this study, we investigated the contribution of color information in predictive power of saliency model by incorporating color to a luminance based model of saliency [1]. We also identified and compared the salient regions of a data set of color videos and same videos in grayscale, through an eye-tracking experiment.

2 Method

2.1 Saliency model

The luminance-based saliency model of Marat et al. [1] draws inspiration from human visual system. The model is consisted of two pathways: static and dynamic. Both pathways are only based on luminance information of visual scene, processed in two steps: The first step simulates some basic pre-processing done by the retina cells through a cascade of three linear filters: a band pass filter for luminance pre processing and two low pass filters for chrominance. Note that we did not model spatially variant resolution of the retina photo receptors. The retina separates the input signal into low and high spatial frequencies that schematically represent the magno- and parvo-cellular outputs of the retina. At second step each signal is decomposed into elementary features by a bank of cortical-like filters. These filters, according to their frequency selectivity, orientation and motion amplitude, provide two luminance-based saliency maps: static map M_{ls} and dynamic map M_{ld}, Figure 1.

The model proposed by Marat et al. is only based on the luminance information. The novelty of our model is to incorporate the color information to compute the saliency map. The early transformation of the Long, Medium and Short wavelength signals, absorbed by cones, provides an opponent-color space in which signals are less correlated [13]. There are several color spaces proposing different combination of cone responses to define the principal components of luminance and opponent colors, red-green (RG) as well as blue-yellow (BY) [14]. The color space proposed by Krauskpof et al. [15] is one of the validated representations to encode visual information where the orthogonal directions, A, $Cr1$ and $Cr2$, represent luminance, chromatic opponent red-green and chromatic opponent yellow-blue respectively. The following equation is used to compute A, $Cr1$ and $Cr2$. In our model we used

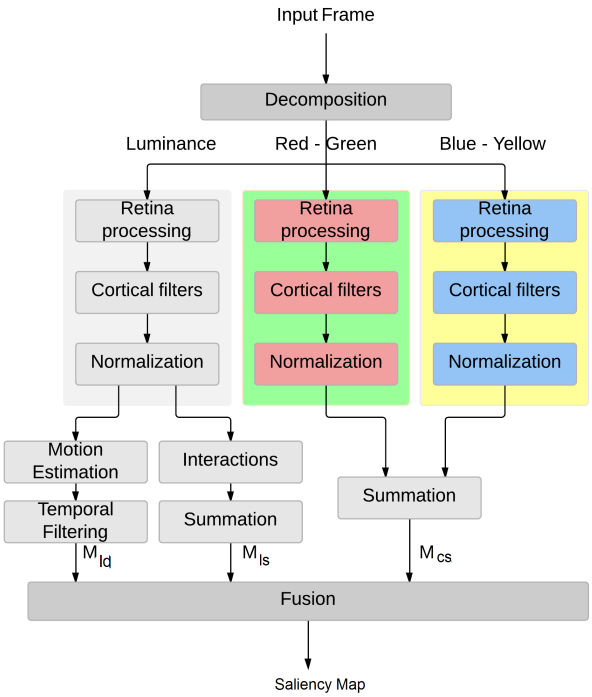

Fig. 1 The spatio-temporal saliency model. M_{ld} is luminance-based dynamic map, M_{ls} and M_{cs} are luminance-based and chrominace-based static maps respectively.

$Cr1$ and $Cr2$ to compute a chrominance saliency map.

$$\begin{pmatrix} A \\ Cr1 \\ Cr2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ -0.5 & -0.5 & 1 \end{pmatrix} \begin{pmatrix} L \\ M \\ S \end{pmatrix}$$

where, L, M and S signals are calculated from tristimulus values of 1931 CIE XYZ color space as follows:

$$\begin{pmatrix} L \\ M \\ S \end{pmatrix} = \begin{pmatrix} 0.4002 & 0.7076 & -0.0808 \\ -0.2263 & 1.1653 & 0.0457 \\ 0 & 0 & 0.9182 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$$

It is known that the human visual system is sensitive to the high spatial frequencies of luminance [16] and the low spatial frequencies of chrominance [17]. The amplitude spectra of the two color-opponent $Cr1$ and $Cr2$ images do not have as many specific orientations as the amplitude spectra of the luminance image [18]. Hence the retinal and cortical processing of chrominance information

is different from luminance information. We integrated to the Marat et al. [1] spatio-temporal saliency model, the chrominance processing steps first introduced by Ho-Phuoc et al. [19]. The retinal processing step of chrominance information starts with low pass filtering illustrated by the contrast sensitivity functions (CSFs) for chrominance information [6]. Following these CSFs, the two color opponents are processed by two low-pass filters. Then the cortical like filters extract the spatial information of $Cr1$ and $Cr2$ color opponents according to 4 orientations (0, 45, 90, and 135 degrees) and 2 spatial frequencies, providing a chrominance static saliency map M_{cs}. Chrominance saliency map M_{cs}, luminance-based static saliency map M_{ls} and dynamic saliency map, M_{ld}, after normalizing, are combined, according to the following equation, to obtain a master spatio-temporal saliency map per video frame. This map predicts the salient regions i.e. the regions that stand out in a visual scene.

$$Saliency\ \ map = \alpha M_{ls} + \beta M_{ld} + M_{cs} + \alpha\beta(M_{ls} \cdot M_{ld})$$

Where, α and β are the max of M_{ls} and skewness of M_{ld} respectively, and $M_{ls} \cdot M_{ld}$ is a pixel to pixel multiplication. Figure 3 shows an example frame and its intermediate and final saliency maps.
In addition, we compared the performance of the model with one of the

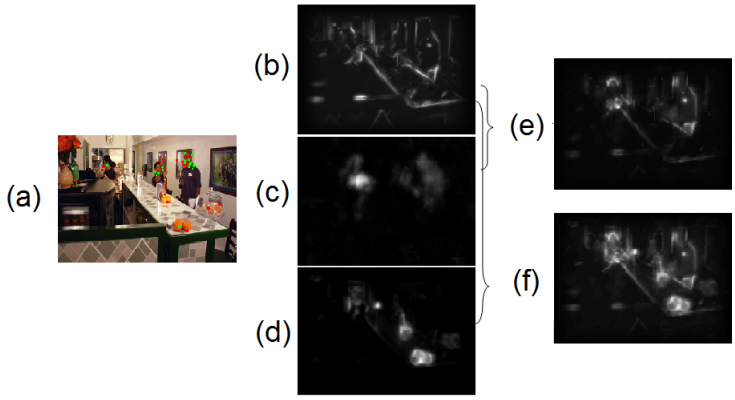

Fig. 2 Saliency maps: (a) An example frame, (b) luminance-based static map M_{ls}, (c) luminance-based dynamic map M_{ld}, (d) chrominance-based static map M_{cs}, (e) fusion of M_{ls} and M_{ld}, (f) fusion of M_{ls}, M_{cs} and M_{ld}.

reference saliency models, Itti and Koch saliency model [20], [4].

GPU implentation The saliency model presented above with static (luminance-based), dynamic (luminance-based) and chrominance pathways is compute-intensive. Rahman et al. [21] have proposed a parallel adaptation of luminance-based pathways onto GPU (`http://www.gipsa-lab.fr/projet/perception/`).

They applied several optimizations subtending to a real-time solution on multi-GPU. We included the parallel adaptation of chrominance pathway to this GPU implementation maintaining the real time solution.

NSS metric A common metric to compare experimental data to computational saliency maps is the Normalized Scanpath Saliency (*NSS*) [3]. We used this metric to compare C and GS eye positions to their equivalent saliency maps. To compute this, first the saliency maps were normalized to zero mean and unit standard deviation. The NSS value of frame k corresponds to averaged saliency values at the locations of eye positions on the normalized saliency map M as shown in the following equation:

$$NSS(k) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\sigma_k} (M(X_i) - \mu_k)$$

where N is the number of the eye positions, $M(X_i)$ is the saliency value of the eye position (X_i), μ_k and σ_k are the mean and standard deviation of the initial saliency map of frame k. A high positive value of NSS indicates that the eye positions are located on the salient regions of the computational saliency map. A NSS value close to zero represents no relation between eye position and the computational saliency map, while a high negative value of NSS means that eye positions were not located on the salient regions of computational saliency map.

2.2 Eye-tracking experiment

To investigate whether the inclusion of color information into saliency model improves its performance, we compared the luminance based and the luminance-chrominance based model to the eye positions of 45 volunteers (25 women and 20 men, range $25 - 39$ years old) recorded while freely viewing videos in two conditions: Color and Grayscale. To simplify, the eye positions recorded when viewing color stimuli and grayscale stimuli are called C positions and GS positions respectively. We also studied the eye positions to determine whether color information influences the eye positions. An Eyelink 1000 from SR research was used to record the eye positions in a pupil tracking mode. The stimuli consisted of 65 short video extracts of 3 to 5 seconds, called video snippets. Video snippets were extracted from various open source color videos. The stimuli measured 640×480 pixels, subtending a visual angle of 25×19 degrees at a fixed viewing distance of 57 cm. The temporal resolution of video snippets was 25 frames per second. The video data set was converted to grayscale using following equation.

$$L = 0.5010 \times R + 0.4911 \times G + 0.0079 \times B \quad (1)$$

The weights of R, G and B channels were calculated according to the experimental display characteristics to fit $V(\lambda)$, the CIE 1931 luminosity function

of standard observer. Display characteristics were obtained by measuring the light emitted from computer-controlled display, using a Photo Research PR650 spectrometer. Figure 3 presents the spectral power distributions of R, G and B channels.

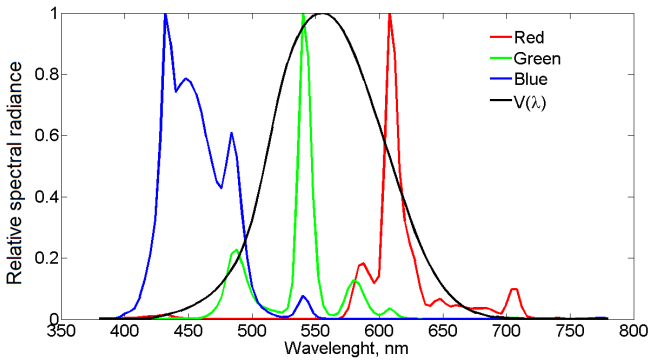

Fig. 3 Spectral power distribution for light emitted by the red, green, and blue phosphors of experimental display and the CIE 1931 luminosity function of standard observer, $V(\lambda)$.

2.3 Eye position analysis metrics

Dispersion. To evaluate variability of eye positions between observers, we used a metric called *dispersion* [1], [22]. Dispersion was calculated separately for each frame for C positions (D_C) and GS positions (D_{GS}). Lower values of dispersion correspond to subjects' eye positions located close to one another, interpreted as high inter-subject consistency.

Clustering. Salient regions of a visual scene can be identified as the locations fixated by a group of subjects at the same moment of observation. These regions can be estimated by clustering the eye positions of different subjects on each frame [23], [24], [25]. Here, we clustered the eye positions to compare the experimental salient regions in color and grayscale conditions using mean-shift clustering method [24]. This method requires a distance parameter to be adjusted. Because the size of video clips was constant, we empirically set this distance to 75 pixels, equal to nearly 3 degrees of visual angle.

3 Results

3.1 Saliency model

First, we studied whether *luminance based saliency model* [1] predicts the eye positions in both conditions with equal efficiency. Then we performed NSS analysis, but using the model of saliency with chrominance. As shown in table 1 color information improves significantly the performance of presented model for both C and GS positions ($GS : t(63) = 4.5, p < 0.01, C : t(63) = 4.86, p < 0.01$), while it improves slightly the performance of the model of Itti and Koch [4].

Table 1 NSS results for Marat et al. model and Itti and Koch saliency model with and without color features.

		Marat		**Itti**	
		luminance	luminance +chrominance	luminance	luminance + chrominance
NSS	**C positions**	0.59	1.18	0.91	0.95
	GS positions	0.60	1.17	0.93	0.97

In addition the GPU implementation of chrominance pathway, similar to luminance-based static pathway results in a speedup of $166\times$ over matlab implementation, while the speedup of dynamic pathway is about $184\times$ over matlab.

3.2 Analysis of eye positions

The dispersion of color eye positions is significantly higher than grayscale (5.1 vs. 4.8, $t(63) = 2,5804, p < 0.01$). This raw result shows that there is more variability between the eye positions of observers when viewing color videos. Yet, a large dispersion might be observed in two different situations: (i) when all observers look at different areas, or (ii) when there are several distant clusters of eye positions. The mean number of clusters on color snippets was significantly higher than grayscale (5.1 vs. 4.8, $t(63) = 2.6, p < 0.01$). The result indicates that the high dispersion value of C positions is not due to the high variability of the eye positions, but it is related to the higher number of regions of interest in color stimuli. However, main clusters were superimposed between C and GS positions. Figure 4 shows the subjects regions of interest on an example frame identified by clustering the C positions and GS positions.

3.3 Conclusion

In the present manuscript, we have compared eye positions recoded while viewing dynamic stimuli in two conditions: color and grayscale. We observed

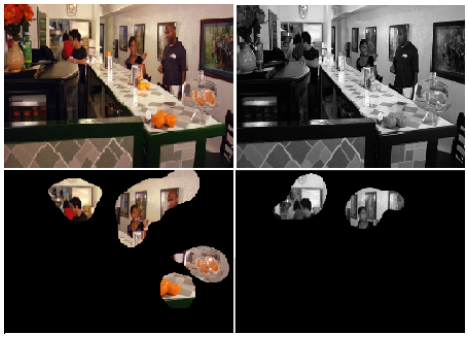

Fig. 4 Example of the regions of interest identified by clustering the eye positions. From left to right, first row: an example frame in color and grayscale. Second row: the corresponding regions of interest of C positions and GS positions.

that the main regions of interest such as faces [26], [27], [28], and moving objects [29], [30], [1] are common in color and grayscale stimuli, but there exist more regions of interest in color stimuli.

We have integrated color information into our bio-inspired saliency model. Results show that indeed color information improves significantly the performance of the model in predicting eye positions for both grayscale and color stimuli while a better prediction power was expected for color stimuli. This might be due to the fact that the major regions of interest are common in both stimuli conditions, but are better enhanced when employing color information. Yet, the incorporation of color information into the model is not optimized. Because the regions of interest are not always located on colored zones, but their neighboring[6]. Whether reinforcement of luminance saliency according to the color information of neighboring zones can improve the predictive power of saliency model remains to be determined.

References

1. Marat, S., Ho Phuoc, T., Granjon, L., Guyader, N., Pellerin, D., Guérin-Dugué, A.: Modelling spatio-temporal saliency to predict gaze direction for short videos. International Journal of Computer Vision **82(3)** (2009) 231–243
2. Connor, C.E., Egeth, H.E., Yantis, S.: Visual attention: bottom-up versus top-down. Current Biology **14** (2004) 850–852
3. Itti, L.: Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. Visual Cognition **12** (2005) 1093 –1123
4. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **20** (1998) 1254–1259

5. Frintrop, S.: VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search. PhD thesis, Fraunhofer Institut Für Autonome Intelliegente Systeme (2006)
6. Le Meur, O., Le Callet, P., Barba, D.: Predicting visual fixations on video based on low-level visual features. Vision Research **47(19)** (2007) 2483–2498
7. Treisman, A.M., Gelade, G.: A feature integration theory of attention. Cognitive Psychology **12** (1980) 97–136
8. Wolfe, J.M., Cave, K.R., Franzel, S.L.: Guided search: An alternative to the feature integration model for visual search. Journal of Experimental Psychology: Human Perception & Performance **15** (1989) 419–433
9. Baddeley, R.J., Tatler, B.W.: High frequency edges (but not contrast) predict where we fixate: A bayesian system identification analysis. Vision Research **46**(18) (2006) 2824–2833
10. Ho-Phuoc, T., Guyader, N., Guérin-Dugué, A.: When viewing natural scenes, do abnormal colors impact on spatial or temporal parameters of eye movements? Journal of Vision **12(2)** (2012) 1–13
11. Frey, H.P., Honey, C., Knig, P.: Whats color got to do with it? the influence of color on visual attention in different categories. Journal of Vision **11(3)** (2008) 1–15
12. Dorr, M., Martinetz, T., Gegenfurtner, K., Barth, E.: Variability of eye movements when viewing dynamic natural scenes. Journal of Vision **10**(10) (2010) 1–17
13. Buchsbaum, G., Gottschalk, A.: Trichromacy, opponent colours coding and optimum colour information transmission in the retina. Proceedings of the Royal Society of London. Series B, **220(1218)** (1983) 89–113
14. Trémeau, A., Fernandez-Maloigne, C., Bonton, P.: Image numérique couleur, de l'acquisition au traitement. Dunod (2004)
15. Krauskopf, J., Williams, D.R., Heeley, D.W.: Cardinal direction of color space. Vision Research **22** (1982) 1123–1131
16. Field, D.J.: Relations between the statistics of natural images and the response properties of cortical cells. J. Opt. Soc. Am. A **4** (1987) 2379–2394
17. Gegenfurtner, K.R.: Cortical mechanisms of colour vision. Nature Reviews Neuroscience **4(7)** (2003) 563–72
18. Beaudot, W.H.A., Mullen, K.T.: Orientation selectivity in luminance and color vision assessed using 2-d bandpass filtered spatial noise. Vision Research **45(6)** (2005) 687–696
19. Ho-Phuoc, T., Guyader, N., Guérin-Dugué, A.: A functional and statistical bottom-up saliency model to reveal the relative contributions of low-level visual guiding factors. Cognitive Computation **2(4)** (2010) 344–359
20. Klab: http://www.klab.caltech.edu/ harel/share/gbvs.php
21. Rahman, A., Houzet, D., Pellerin, D., Marat, S., Guyader, N.: Parallel implementation of a spatio-temporal visual saliency model. Real-Time Image Processing **6(1)** (2010) 3–14
22. Salvucci, D., Goldberg, J.H.: Identifying fixations and saccades in eye-tracking protocols. Symposium on Eye tracking research applications **469**(1) (2000) 71–78
23. Follet, B., Le Meur, O., Baccino, T.: New insights on ambient and focal visual fixations using an automatic classification algorithm. iPerception **2(6)** (2011) 592–610
24. Santella, A., DeCarlo, D.: Robust clustering of eye movement recordings for quantification of visual interest. In: Eye Tracking Research and Applications (ETRA) Symposium. (2004)

25. Coutrot, A., Guyader, N., Ionescu, G., Caplier, A.: Influence of soundtrack on eye movements during video exploration. Journal of Eye Movement Research **5**(4) (2012) 1–10
26. Rahman, A., Houzet, D., Pellerin, D.: Influence of number, location and size of faces on gaze in video. Journal of Eye Movement Research **7(2)** (2014) 1–11
27. Marat, S., Rahman, A., Pellerin, D., Guyader, N., Houzet, D.: Improving visual saliency by adding face feature map and center bias. Cognitive Computation **5(1)** (2013) 63–75
28. Rousselet, G.A., Macé, M.J.M., Fabre-Thorpe, M.: Is it an animal? is it a human face? fast processing in upright and inverted natural scenes. J. Vision **3**(6) (2003) 440–55
29. Itti, L., Baldi, P.: Bayesian surprise attracts human attention. Vision Research **49(10)** (2009) 1295–1306
30. Mital, P.K., Smith, T.J., Hill, R.L., Henderson, J.M.: Clustering of gaze during dynamic scene viewing is predicted by motion. Cognitive Computation **3(1)** (2010) 5–24